DOCUMENT RESUME

ED 217 074 TM 820 328

AUTHOR Smith, Laura Spooner; And Others
TITLE Characteristics of Student Writing Competence: An Investigation of Alternative Scoring Systems.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-RR-134
PUB DATE 80
NOTE 36p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College English; *College Freshmen; Comparative Analysis; Higher Education; High School Seniors; *Scoring; *Screening Tests; *Student Placement; *Writing Skills

ABSTRACT

Three alternative methods for placing post-secondary students into freshman English or remedial writing classes are compared. The study contrasted: (1) a proposed system-wide test combining multiple choice and essay scores; (2) the holistic essay scoring procedures used at separate university campuses; and (3) an analytic scoring rubric developed at a university-based research center. The comparability of scores obtained from the three methods and the placement decisions they implied are examined. High school seniors from two university campuses took an experimental version of the proposed system-wide placement examination. Relationships were low among scores from the different testing methods, and substantially different proportions of students were classified as masters or non-masters. These findings were interpreted as evidence that "good" writing does not consistently emerge, regardless of the test used and that systematic selection of placement measures requires detailed scrutiny of the reliability and validity of placement standards, scoring criteria and their emphasis on essay features. (Author/GK)

# CHARACTERISTICS OF STUDENT WRITING COMPETENCE: AN INVESTIGATION OF ALTERNATIVE SCORING SYSTEMS

Laura Spooner Smith, Lynn Winters
Edys S. Quellmalz, and Eva L. Baker



CSE Report No. 134
1980



CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

# Effects of Alternate Scoring Options on the Classification of Entering Freshman Writing Competencies

## Abstract

This study compared three alternative methods for placing post-secondary students into freshman English or remedial writing classes. The study contrasted: 1) a proposed system-wide test combining multiple choice and essay scores; 2) the holistic essay scoring procedures used at separate university campuses; 3) an analytic scoring rubric developed at a university-based research center. The study examined the comparability of scores obtained from the three methods and the placement decisions they implied.

Three hundred eight high school seniors from two university campuses took an experimental version of the proposed system-wide placement examination. Generally, relationships were low among scores from the different testing methods, and substantially different proportions of students were classified as masters or non-masters. These findings were interpreted as evidence that "good" writing does not consistently emerge, regardless of the test used and that systematic selection of placement measures requires detailed scrutiny of the reliability and validity of placement standards, scoring criteria and their emphasis on essay features.

Among the many criticisms of the quality of public education, complaints about students' inability to write prose lead the pack. At the time of college admission when students need to be assigned to beginning English courses, writing deficiencies become especially salient. At entrance to college, students may be assigned to college-level beginning English courses, or with greater frequency, may be placed in a special course designed to remedy composition problems and to prepare for regular college level work. This initial placement decision is made through different means. Some schools base their decision solely on student verbal scores on a college entrance examination. Others require that all students take a special placement examination. These examinations may vary in their development history (locally prepared or commercially published), definition of writing (narrative or expository prose), format (multiple choice or essay production), and manner by which the passing score is determined. An ideal and experimentally clean way to make choices among such alternatives would involve the systematic variation of some of these variables to determine which procedures provide the least mistaken estimate of students' writing ability. In fact, admission is a serious business and little experimental "fooling" with the system is tolerated in real colleges and universities, even for the promised benefit of improved decisions.

This study, however, is an attempt to contrast alternative assessment methods in actual placement testing. Its practical impetus grew from specific requirements in the higher education system in California. As

background, California has two, state-wide university systems: The University of California (UC) and the California State University and Colleges (CSUC). Although the systems are designed to attract different levels of students (at UC, the top 12½% statewide and at CSUC, the top 33%) students may transfer from system to system or to different campuses within the same system. CSUC consists of 19 campuses, and to standardize requirements among campuses, a committee of faculty cooperated with the Educational Testing Service (ETS) to develop a system-wide test of English composition placement, the English Placement Test (EPT). The UC system of nine campuses operates so that each campus' unique placement test (called the Subject A examination) is honored by the other campuses. Since CSUC students often wish to transfer to UC schools, a study group made up of faculty from both systems was appointed to review the need for common writing placement procedure for all UC and CSUC campuses. The use of the English Placement Test was suggested by the CSUC representatives.

The problem in its most simple form is whether the EPT would provide the same quality of information thought to be obtained through the existing procedures at UC campuses. Could a test designed for a population consisting of the top one-third of students operate efficiently for the top 12½%?

Embedded in this problem are a number of serious issues related to the teaching and testing of writing. For a start, few agree on the definition of writing competence itself. A common, but operationally vague desire is that students ought to write well enough to succeed in other college courses, as if success were an unidimensional phenomenon. In fact, Smith (1975) demonstrated that requirements for success vary from college

specialization to specialization. Definitions of competence may focus on particular features of writing, such as structural or grammatical elements. In other views, acceptable mechanics are a minimum, but emphasis is given, in addition, to the quality of thought or to the logic and clarity of the communication.

A second issue running through this study is the form of student response used to make the decision. Some tests of writing rely heavily on "indirect" measurement, where performance on multiple choice tests is used to "predict" writing achievement. These tests are justified along these connected lines of argument. First, the correlation coefficients of written essays and multiple choice tests are high enough that the "validity" of the objective test should not be challenged. The tests are functionally thought to measure the "same thing" (Godshalk, Swineford, & Coffman, 1966; Breland & Braucher, 1977). Given this equivalence, efficiency favors choosing the least expensive method, and objective tests are easier and cheaper to administer and score. The scoring argument is bolstered by the well-known differences in raters' judgments of essays, that is, the matter of scorer unreliability.

Proponents of collecting writing samples from students argue that the cognitive requirements of creating essays and answering a series of multiple choice tests differ markedly from one another, and that no amount of statistical modelling can actually equate writing with choosing the right answer (Spooner-Smith, 1978; Quellmalz & Capell, 1979). Further criticisms of rater unreliability are countered by the results of good training procedures. However, the cost issue remains, cast by these advocates as a choice between cheap, irrelevant information or more

costly, valid data.

A third issue applies to any definition or format for the assessment of student writing competence: How are standards of passing or failing set? Does the standard treat equally the two forms of potential misclassification, competent students who "fail" and incompetent students who "pass"? Is there a policy that the benefit of the doubt goes to the student? Does the system so value its definition of writing that it wishes to be conservative about who gets to enter college English courses?

A last, but critical issue arises for those who have opted for the collection of essay responses. Not only questioned are the number, type, and length of responses necessary for accurate judgment, but also heated disagreement occurs over the best scoring procedures. The choices are between holistic scoring, which gives an overall estimate of the essay, and analytic scoring which provides subscores for particular characteristics of the writing. Again, the conflict is between cost, where holistic scoring takes approximately 2/3 the time of analytic scoring, and precision of information, where analytic scores provide diagnosis of deficient performance. Strong advocates for holistic scoring cite its economy (Godshalk, et al., 1966; Alloway, 1978; Powills, Bowers & Conlan, 1979). However, feature analyses of good and poor papers point to the distinct differences in their content and structure (See Cooper, Cherry, Gerber, Fleischer, Copley, & Sartisky, 1979), and advocates of analytic ratings argue for the use of such information in determining instructional policy for remediation (Quellmalz, 1980).

With contention as a backdrop, then, the practical problem of choosing

a "good" placement procedure for UC was studied. Staff at a university-based research center proposed research to compare three alternative methods for making the placement decision: the use of the English Placement Test (EPT) (consisting of an essay and multiple choice scales) proposed by the CSUC staff; the placement procedure (Subject A examinations) in use at each of the two UC campuses; an analytic essay rating scale developed by the research center in the course of its studies of writing (the CSE scale). Two simple questions were formulated to guide this study:

1. How comparable are the scores students receive from each form of writing assessment?
2. Would the methods sort students in competent and incompetent groups in the same way?

# METHODS

## Overview

Each of two UC campuses agreed to participate in the study. Instead of requiring their own Subject A examination, each campus administered the EPT examination to a sample of students participating in regular placement examinations. The EPT essay was first scored by ETS, rescored at each campus using campus scoring procedures (both campuses used holistic rating procedures), and then the essays were sent to the research center for re-rating according to the CSE analytic scheme. Actual placement decisions for each student were made on the basis of the campus interpretation of ETS scores.

## Subjects

Three hundred eight high school seniors were required to take the experimental version of the placement examination at either of two UC campuses. A placement test for writing was a regular requirement for students scoring between 450 and 600 on the College Entrance Examination Board (CEEB) test.

## Instruments

### The English Placement Test

The EPT was developed by the Educational Testing Service in collaboration with CSUC as a placement tool for first-year English classes in the CSUC system. The EPT requires students to write one 45-minute essay and to complete a 90-minute multiple choice section covering three skill areas: reading, sentence construction, and logic and organization. The reading

section asks students to identify main ideas and to interpret ideas in short reading passages. The sentence construction test items require students to recognize arrangements of sentence elements that "express meaning clearly and correctly." The logic and organization section contains a variety of item types intended to measure students' ability to "see relationships between words." For example, some items require students to arrange words into categories; other items involve identifying sentences to begin, end, or support a given paragraph. Still other items intend to measure the students' ability to distinguish between fact and opinion. The objective part of the EPT counts 75% of the total.

Essay topic. The essay direction required students to write a 45-minute essay on a topic eliciting narrative/descriptive writing. The topic of this administration called for students to write about "a real or an apparent change that had occurred in someone they knew."

EPT essay criteria. The EPT scoring scale is a six-point holistic essay scale divided into two parts--"upper half papers" and "lower half papers." Raters are instructed to read each paper through quickly and assign an overall rating based on how well the essay addressed itself to all aspects of the question (topic), how well the essay is organized, and how well it demonstrates writing quality. Aspects of writing quality mentioned in the rubric are syntax and diction. Papers that do not respond to, argue or avoid the question are scored zero. The EPT was studied for content validity, as reported by Breland and Ragosa (1976). Unfortunately, no results were available.

UC Campus 1: holistic essay criteria

Campus 1 employed a six-point holistic scale which permits readers to assign a plus or minus to each point on the scale (1=high, 6=low). The rubric directs raters' attention to the thesis statement and its development, sentence structure, word choice, and a detailed list of "mechanics" features. Additionally, each point on the scale corresponds to a placement decision. For example, scores of one, two or three indicate that the student is prepared to take a regular freshman composition course, while a score of four through six indicates that the student should be placed in one of a series of increasingly remedial English classes. Campus 1 typically employs a one-hour placement examination.

Campus 2: holistic essay criteria

A six-point holistic rating scale was also employed by Campus 2 (1=low, 6=high). The rubric emphasizes fluency and mechanics, although reference is made to the logic and organization of the writing scale. In its normal placement examination, two one-hour essays are produced by each student at Campus 2.

CSE analytic essay criteria

Unlike the three holistic approaches of the other rating procedures, the CSE essay scoring provides an analytic rating of each essay (Quellmalz, 1979). The analytic rubric derived from other scales used for narrative discourse and from texts and tests in composition and rhetoric (Pitts, 1978). The scale presents carefully explicated criteria developed for domain-referenced narrative writing tasks. Scale criteria require refer-

ence to observable features in an essay, unlike many rating rubrics which include more subjective, affective judgments. The scale consists of five subscales, each with a range of four points. Based on studies suggesting that holistic and analytic ratings provide distinct information about student writing, the scale calls for both holistic and analytic ratings (Winters, 1978). The first subscale, General Impression, directs raters to read the paper quickly first and to rate it according to their global judgments of its quality as an example of narration. The remaining four subscales attend to the following components of the writing: focus, organization, support, and mechanics. The scoring rubric for the scale contains a detailed description of essay features associated with each of the four levels of quality within each of the subscales.

### Archival student information

In addition to the three scores generated by the rescoring of the required placement exam, Scholastic Aptitude Test (SAT) verbal scores, College Entrance Examination Board (CEEB) scores, High School English course grades and grade point averages were also available for students.

## Procedures

### Administration

Students who came to the required UC placement examination were divided, as they arrived, into groups taking the regular or the experimental EPT administration. Students in the study were placed in the same room and not exposed to the usual campus procedure. The entire EPT was administered according to the publisher's directions. This process was repeated

on each of the two UC campuses in the study.

EPT Scoring Procedure

The essays rated by the EPT procedures were graded at the same time as a larger pool of essays from all CSUC campuses (n=6,293). Twenty-seven raters were trained in a three and one-half hour training session to assign scores according to the EPT rubric. Each essay was read by two readers and the final score assigned to an essay was the sum of the two scores. As the EPT rubric was a six-point scale, essay scores ranged from one to twelve. Papers with scores differing by two or more points and all papers that received a zero score from one reader but a non-zero score from the other reader were read by a third reader. The total essay score in these adjudicated cases was the sum of the two most congruent scores. EPT reported that the majority of discrepant scores occurred in the three to five score range.

Rater agreement was calculated by a correlation coefficient summarizing the amount of agreement between the first and second scores assigned to a paper, rather than of the amount of agreement between particular rater pairs. The correlation coefficient reported for 5,756 papers was .59.

CSE Rating Procedures

The combined set of 308 essays was rescored at the research center using the CSE Factual Narrative Scale II. Four raters, English instructors, were hired to read the essays. All of the raters had previous experience in the systematic rating of student essays, and two of the four raters had used the particular scale in previous studies.

CSE rater training procedures were similar to those employed by Spooner-Smith (1978) and Winters (1978). Approximately four hours were devoted to

review, rating and discussion of 30 sample essays on the essay topic. At the conclusion of the training session, rater agreement coefficients were computed for each of the subscores and the total scale in order to determine whether training should be continued. Alphas ranged from .86 to .92 (based on four ratings per paper), and generalizability coefficents ranged from .59 to .87. As a result, readers reread and discussed the pilot test papers again for the one subscale with low reliability, focus, before reading the actual "experimental" essays. Papers were randomly assigned to raters.

Campus 1: rating procedure

Six teaching assistants experienced in teaching basic writing rated the Campus 1 essays returned by ETS. The Campus 1 scale, based primarily on a tally of mechanical errors, was used to assign essay scores. Each paper was read by one reader; raters were department teaching assistants and were given no additional formal training.

Campus 2: rating procedure

Campus 2 papers were read by seven raters, all composition instructors. The raters had previous experience in rating placement essays for the English department, so only about one and a half hours were devoted to rater training. During this session, raters read and discussed essays on topics analogous to the EPT topic and assigned scores according to the Campus 2 writing exam scale.

Each paper was read by two raters; the final score was the sum of the two ratings. Papers discrepant by two or more points were read by a third reader and the discrepancy resolved in the same manner as were discrepan-

cies in the EPT scoring procedure. Campus 2 calculated no interrater reliabilities.

## RESULTS

### Comparability of Assessment Procedures

The first section of results addresses the comparability of the three alternative measures and includes internal analyses of each (see Table 1). The EPT and CSE scores will be treated first because they each provide subscales. Consider the EPT analyses. The most dramatic

- - - - - - - - - - - - - - - -
Insert Tables 1 & 2 about here
- - - - - - - - - - - - - - - -

findings surround the relationship of the objective EPT subscales and the essay score (see Table 2). Each of three subscales strongly correlates with one another, a fact which suggests that they may provide redundant information. These subscales, taken individually or combined into an "objective" composite relate only moderately with the EPT essay score analyses (ranges of $r$ between .25 and .30).

The CSE scale analysis addresses the relationship of the four analytic subscores, the total of these scores, and the General Impression, "holistic" score for each essay (see Table 3). The relatively low correlations sug-

- - - - - - - - - - - - - - -
Insert Table 3 about here
- - - - - - - - - - - - - - -

gest that the particular subscales are, in fact, identifying separate skill

Table 1

Means and Standard Deviations

| | Possible | n | $\bar{X}$ | s.d. | n | $\bar{X}$ | s.d. |
|---|---|---|---|---|---|---|---|
| EPT TOTAL | 180 | 104 | 152.38 | 6.44 | 201 | 154.17 | 3.84 |
| EPT ESSAY | 12 | 104 | 7.03 | 1.38 | 201 | 7.37 | 1.58 |
| EPT OBJECTIVE SCALES | | | | | | | |
| Reading | 180 | 104 | 153.04 | 8.81 | 201 | 154.20 | 12.10 |
| Sentence construction | 180 | 104 | 154.72 | 7.35 | 201 | 156.01 | 11.89 |
| Logic and organization | 180 | 104 | 153.16 | 7.89 | 201 | 154.01 | 11.95 |
| Composition | 180 | 104 | 152.03 | 6.13 | 201 | 153.09 | 11.53 |
| Total Objective Score | 540 | 104 | 460.92 | 22.11 | 201 | 464.21 | 35.00 |
| CAMPUS SCORING | | 103 | 2.93 | 1.26 | 201 | 6.61 | 2.02 |
| CSE SUBSCALES | | | | | | | |
| General Impression | 4 | 69 | 1.52 | .71 | 148 | 1.79 | .78 |
| Focus | 4 | 69 | 1.80 | .56 | 148 | 1.98 | .55 |
| Organization | 4 | 69 | 1.70 | .63 | 148 | 2.00 | .70 |
| Support | 4 | 69 | 1.88 | .67 | 148 | 2.09 | .69 |
| Mechanics | 4 | 69 | 1.91 | .53 | 148 | 2.35 | .62 |
| Total | 20 | 69 | 8.81 | 2.35 | 148 | 10.17 | 2.52 |

TABLE 2

Internal Characteristics of EPT and CSE Assessment

English Placement Test

| EPT | Essay | Reading | Sentence construction | Logic | Composition | Objective | Total |
|---|---|---|---|---|---|---|---|
| Essay | | | | | | | |
| Reading | .27 | | | | | | |
| Sentence construction | .28 | .68 | | | | | |
| Logic | .25 | .71 | .62 | | | | |
| Composition | .71 | .70 | .79 | .78 | | | |
| Objective Total | .30 | .91 | .86 | .88 | .85 | | |
| Total | .62 | .85 | .81 | .81 | .97 | .93 | |

N= 308

TABLE 3

Center for the Study of Evaluation analytic scale

| CSE Scale | General Impression | Focus | Organization | Support | Mechanics | Total |
|---|---|---|---|---|---|---|
| General Impression | | | | | | |
| Focus | .47 | | | | | |
| Organization | .75 | .47 | | | | |
| Support | .48 | .46 | .55 | | | |
| Mechanics | .46 | .41 | .32 | .28 | | |
| Total | .85 | .72 | .83 | .73 | .65 | |

N = 217

components. The correlation of .85 for the General Impression and the total of the subscales suggests that directing one's attention to four particular features of writing nonetheless produces values consistent with an overall holistic view.*

The comparison between features assessed by the EPT and the CSE indicators more directly addresses the question of assessment comparability (see Table 4).

- - - - - - - - - - - - - -
Insert Table 4 about here
- - - - - - - - - - - - - -

The essay scores derived from EPT and CSE scoring suggest that only a moderate amount of overlap exists in the scoring rubrics. The holistic ratings between the CSE General Impression and EPT essay correlate in the mid-ranges; however, the component skills measured by the CSE analytic dimensions and the EPT subscales diverge dramatically. For instance, "organization" is assessed by both EPT and CSE scores, yet the correlation between subscales is only .12. Sentence construction on the EPT and mechanics on the CSE subscale, apparently comparable dimensions, correlate .29. Clearly, the format of the EPT subscale responses (objective tests) assesses a different capacity than the CSE subscale rating of the essay.

Comparisons were also made among the EPT scores, CSE scores, and the UC campus holistic scoring procedures. In Table 5, the first column pre-

- - - - - - - - - - - - - -
Insert Table 5 about here
- - - - - - - - - - - - - -

*In fact, the holistic score is undoubtedly contaminated by the raters' use of the analytic rating scales, after the first paper, that is.

TABLE 4

Cross-Correlations Between EPT and CSE Subscales

Campuses Combined

| | CSE | | | | | |
|---|---|---|---|---|---|---|
| EPT | General Impression | Focus | Organization | Support | Mechanics | Total |
| Essay | .46 | .46 | .41 | .42 | .38 | .56 |
| Reading | .17 | .15 | .14 | .16 | .27 | .23 |
| Sentence construction | .18 | .18 | .16 | .15 | .29 | .25 |
| Logic & organization | .14 | .20 | .12 | .11 | .23 | .21 |
| Composition | .36 | .39 | .32 | .31 | .40 | .4 |
| Objective test | .19 | .20 | .16 | .16 | .30 | .27 |
| Total | .39 | .33 | .28 | .28 | .39 | .42 |

TABLE 5

Correlation of Placement Test Scores from EPT, Campus 1 Campus 2, and CSE

| | EPT essay | EPT objective | Campus 1 | Campus 2 | CSE |
|---|---|---|---|---|---|
| EPT essay | | | | | |
| EPT objective | .30 | | | | |
| Campus 1 | .60 | .53 | | | |
| Campus 2 | .25 | .08 | * | | |
| CSE | .40 | .27 | .48 | .12 | |

*Campus 1 and 2 scored only their own students' essays.

sents the simplest contrasts. The EPT correlates at the .40 level with the CSE total. The holistic scoring procedures at the UC campuses results in discrepant relationships (at Campus 1, r=.60, and at Campus 2, r=.25). A low risk conclusion is that "holistic" ratings (as used at each campus and for the EPT rating) mean different things. In any case, inferences about the stability of these relationships is certainly weakened by the relatively low inter-rater reliability reported for the EPT ratings, the lack of reliability estimates for the UC efforts, and the potential for error inherent in the single rating procedure used at Campus 1. Yet, even if these ratings were reliable, the conclusion from these data would be that raters using different systems operationalize writing in very different ways.

Relationship of assessment procedure and archival information

Table 6 presents descriptive statistics for archival data by campus and Table 7 displays the correlations among different writing assessment methods and other writing-related archival data often used in placement decisions. Making inferences from such spotty results is dangerous; however, the most consistent relationships are among the College Entrance,

- - - - - - - - - - - - - - - - -
Insert Tables 6 & 7 about here
- - - - - - - - - - - - - - - - -

Scholastic Aptitude, and English Placement Tests. While this relationship may result from connections between underlying abilities (for instance, comprehension ability is assessed on all three measures), one might argue that the fact that these tests originate from the same publisher, using

TABLE 6

Means and Standard Deviations
for Archival Data* by Campus

| | Campus 1 | | | Campus 2 | | |
|---|---|---|---|---|---|---|
| | N | $\bar{X}$ | s.d. | N | $\bar{X}$ | s.d. |
| High School English Grades | 61 | 3.68 | .34 | 187 | 3.70 | .35 |
| High School Grade Point Average | 90 | 3.68 | .28 | 161 | 3.68 | .28 |
| College Entrance Examination Board Scholastic Aptitude Test (Verbal) | 90 | 478 | 79 | 184 | 509 | 63 |
| | 90 | 492 | 87 | 180 | 510 | 83 |

TABLE 7

Correlations Between Alternative Placement Scores and Other Predictors of College English Performance

| | College Entrance Examination Board | Scholastic Aptitude Test (Verbal) | High School English | High School Grade Point Average |
|---|---|---|---|---|
| EPT total score | | | | |
| Campus 1 | .54 | .59 | .14 | .20 |
| Campus 2 | .66 | .64 | .32 | .31 |
| Combined | .62 | .62 | .19 | .25 |
| CSE total essay score | | | | |
| Campus 1 | .26 | .25 | -.04 | -.01 |
| Campus 2 | .29 | -.01 | .05 | .23 |
| Combined | .32 | .21 | .00 | .07 |
| Campus essay score | | | | |
| Campus 1 | .22 | .23 | .07 | .01 |
| Campus 2 | .50 | .31 | .20 | .31 |

supposedly similar test development technology, may be as plausible a link among them.

More disheartening, however, is the lack of relationship among writing indices and high school and English grade point average. Although range restriction definitely must be considered (all students have a 3.2 minimum grade point average to qualify for UC admission), one would still hope that the grades of these students, drawn as they were from the middle of the CEEB distribution (450-600 scores), might support the validity of the measures. One gloomy view is that high school performance, as measured by grades, does not include much writing competence. Research on the amount of actual precollegiate writing required of students supports this analysis (Pitts, 1978).

A related question is the amount of performance that can be inferred to be a specific skill and the amount inferred to be general ability or perhaps general information. The relatively higher values for the Campus 2 procedures may be explained as general ability. This explanation is especially interesting in the light of the weak categories in the scoring rubric, and the form of rater training. When no need exists for identification and operational statement of criteria in order to achieve set levels of agreement among raters, it is reasonable to infer that the writers' general ability rather than specific writing skill is detected by the rating.

Alternative placement decisions using three assessment models

To compare the utility of the three methods in view of different

standards for pass and fail, two analyses were performed: 1) the pass score was set at the mean of the scores from the experimental UC distribution; 2) the cut score set according to present or recommended practice. The best approach for identifying the optimal placement of such standards would naturally depend upon developing an adequate estimate of "future success" in college writing, and working back from it, to identify the minimum requirements for competency. In the absence of such a refined external criterion, the alternative placement analyses shed light on the differences in decisions made by the various assessment approaches.

Group analyses

At the group level of analysis, Table 8 displays percentages of students who would be placed in remedial classes if cut-off scores were 1) set at the mean of the UC sample for each of the three methods or 2) set at the recommended or regularly used standard. When the cut-off for the EPT essay is set at the UC mean (a customary ETS procedure), 54% of

- - - - - - - - - - - - -

Insert Table 8 about here

- - - - - - - - - - - - -

UC students would be required to take remedial English. If the EPT cut-off score were set at the average of the CSUC population, only 26% of the UC sample would be placed into remedial English. This contrast reflects the differences in populations in the two university systems and suggests that if the EPT essay (and its cut-off) were adopted directly from CSUC, then the standard of writing expected at UC would drop. The CSE scale would place 61% of UC students in the remedial course, with either the average or a substantively set criterion score of 10.

TABLE 8

Percent of Students Placed in Subject A
by the Three Scoring Systems

| | When cut-off scores = UC mean | | | When cut-off scores = those previously used | | |
|---|---|---|---|---|---|---|
| Combined campuses | N | Score | Remedial English | N | Score | Remedial English |
| EPT essay | 304 | < 7.28 | 54 | 304 | ≤ 6 | 26 |
| EPT total | 304 | <153.62 | 48 | 304 | ≤150 | 18 |
| CSE total | 235 | < 9.83 | 61 | 235 | ≤ 10 | 61 |
| Campus 1 | | | | | | |
| Campus rubric | 103 | < 2.93 | 49 | 104 | ≤ 4 | 31 |
| EPT essay | 103 | < 7.03 | 63 | 104 | ≤ 6 | 34 |
| EPT total | 103 | <152.38 | 35 | 104 | ≤150 | 20 |
| CSE total | 71 | < 8.61 | 51 | 71 | ≤ 10 | 79 |
| Campus 2 | | | | | | |
| Campus rubric | 201 | < 6.61 | 40 | 200 | ≤ 7 | 40 |
| EPT essay | 201 | < 7.37 | 50 | 200 | ≤ 6 | 23 |
| EPT total | 201 | <154.27 | 43 | 200 | ≤150 | 14 |
| CSE total | 164 | < 10.35 | 53 | 164 | ≤ 10 | 53 |

Contrasts in performance between the two UC campuses demonstrate that Campus 2 apparently draws from a somewhat more proficient population of writers than Campus 1.

Individual placement decisions

Different predictions can be made about the placement of any individual student under the three assessment methods (see Table 9). Numbers in the "off" diagonal represent students who would pass under one system

- - - - - - - - - - - - - -
Insert Table 9 about here
- - - - - - - - - - - - - -

and fail according to another (taking pairs of procedures one at a time for each campus). For example, at Campus 1, if the pass score were set at the CSUC mean, 30% of the students who pass the EPT essay would fail using the regular standards of the campus, and 57% would fail using the CSE scale. Placement discrepancies between CSE and Campus 1 procedures are greater than between Campus 1 and the EPT decisions. Campus 2 placement decisions similarly demonstrate discrepancies, but with different details. For instance, in comparing the CSE with Campus 2 standards, one can see that 36% of the students would pass in one system and fail in the other. However, the degree of difficulty (as judged by the percentages passing and failing in either system) shows rough equivalence. Thus, in the case of the Campus 2-CSE comparison, it is the defintion of writing competency that accounts for differences in placement rather than "difficulty" of the measure.

TABLE 9

Comparison of Placements When Essay Cut-off Scores
Are Set at Previously Employed Standards

Campus 1 rubric

| EPT essay rubric | Pass ≤3 | Fail ≥4 | |
|---|---|---|---|
| Pass ≥7 | 37 | 31 (30%) | 68 |
| Fail ≥6 | 31 (30%) | 4 | 35 |
| | 68 | 35 | 103 |

Campus 2 rubric

| EPT essay rubric | Pass ≥8 | Fail ≤7 | |
|---|---|---|---|
| Pass ≥7 | 79 | 76 (38%) | 155 |
| Fail ≤6 | 9 (5%) | 36 | 45 |
| | 88 | 112 | 200 |

CSE rubric

| Campus 1 rubric | Pass ≤11 | Fail ≥10 | |
|---|---|---|---|
| Pass ≤3 | 5 | 40 (57%) | 45 |
| Fail ≥4 | 10 (14%) | 15 | 25 |
| | 15 | 55 | 70 |

CSE rubric

| Campus 2 rubric | Pass ≥11 | Fail ≤10 | |
|---|---|---|---|
| Pass ≥8 | 47 | 29 (18%) | 76 |
| Fail ≤7 | 30 (18%) | 58 | 88 |
| | 77 | 87 | 164 |

CSE rubric

| EPT essay rubric | Pass ≥11 | Fail ≤10 | |
|---|---|---|---|
| Pass ≥7 | 15 | 33 (47%) | 48 |
| Fail ≤6 | 0 (0%) | 23 | 23 |
| | 15 | 56 | 71 |

CSE rubric

| EPT essay rubric | Pass ≥11 | Fail ≤10 | |
|---|---|---|---|
| Pass ≥7 | 70 | 57 (35%) | 127 |
| Fail ≤6 | 7 (4%) | 29 | 36 |
| | 77 | 86 | 163 |

## DISCUSSION

The findings of the study dramatize the dilemma facing multi-site educational systems attempting to establish uniform writing competency testing. The question is whether newly proposed placement method B is better than extant placement method A, and the answer is, in this case unfortunately, "It depends." It depends on what you are looking for and what evidence will convince you that you have found it. This study underscores the fact that writing is not an undifferentiated skill construct and that different tests may measure or emphasize very different aspects of the writing competency domain.

The questions guiding this study structured information about the consequences of using different assessment methods: 1) Are descriptions of student writing competence provided by the proposed placement exam comparable to campus methods in use or to an analytic essay scoring scheme? and 2) Do alternative placement methods result in the same placement decisions? The answer to both of these questions is, basically, "No."

The data indicate that descriptions of a student's writing competence derived from the three alternative measures, the EPT (essay and objective tests), the local campus rubrics, and the CSE essay scale differ considerably. These differences are indicated by the generally low correlations among the placement methods and other writing-related indices, and, most importantly, by the discrepant classification of the same student as master or non-master. These empirical analyses suggest a need to return to a logical and psychological analysis of the content of the three measurement

approaches as they relate to what is meant by writing competence.

The low or moderate correlations of the ratings generated by the EPT, UC campus and CSE rubrics imply that the criteria in these scales emphasize different essay features. A look at the content of the rubrics confirms these differences. Even when nominally similar methods were used, empirical differences were found. For instance, both the EPT and Campus 2 rubrics were applications of the ETS holistic scoring procedures applied in large scale writing assessments (Conlan, 1976; Alloway, 1978; Powills, et al., 1979). Yet the same basic approach results in clearly different specifications and applications of criteria by different sets of raters. These results, at minimum, challenge the stability and validity of holistic scoring for placement and competency decisions, where it is critical that consistent criteria be applied fairly to all students.

Our data illustrate that, contrary to folklore, competent writing does not "surface" apart from the details of the rating scheme. The view of writing competency reflected in any rating procedure vastly influences what happens to students. The results of this study were presaged by earlier work. In a study of the effects of alternative response criteria in holistic, analytic and quantitative rating schemes, Winters (1978) also found that the scales differentially profiled the same set of essays and characterized students as masters or non-masters. Furthermore, she reported that imprecisely worded criteria were refined and clarified by raters during training, and she hypothesized that a new set of raters would refine and apply the criteria differently.

This study suggests that the design of writing placement assessments require detailed and systematic consideration of a range of test development issues. Methodology for designing domain-referenced tests (DRT) in general (Hively, 1974; Baker, 1974; Popham, 1978, 1980) and for domain-referenced writing assessment in particular (Quellmalz, 1978, 1980; Baker & Quellmalz, 1979) may provide a useful approach to developing or selecting writing assessments. Such methods begin with a detailed definition of desired writing competencies and then require precise domain specifications for the rhetorical features of the writing task, explicit criteria in the rating scale, and reliable procedures for using the scale. These specifications permit examinations of the planned placement test by subject matter and testing experts prior to the test administration. For example, screening of the task structure and scoring procedures in this study might have resulted in changing the essay task from a narrative one to an expository task more representative of the type of writing required in college courses. Examination of the planned scoring methods might have resulted in the calculation of interrater reliability for Campus 2 and for the scoring of placement essays by more than one rater for Campus 1.

The design of the domain of task and scoring features for a particular placement test also can provide a blueprint for guiding development of comparable, parallel writing tasks, rating criteria and rating procedures, assuring the fairness of decisions from occasion to occasion and site to site. In the ideal case, evidence should indicate that the placement test discriminates between surviving and floundering college writers. This study emphasizes the need for a systematic approach to selecting or developing

writing competency tests. Perhaps through domain-referenced testing methods and continuing longitudinal research on writing assessment problems, we can improve the confidence we place in decisions about writing ability.

## References


Alloway, J. E. Some ways of establishing criteria for assessing writing performance from the perspective of the test developer. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Baker, E. L., & Quellmalz, E. S. Results of pilot studies. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (OB-NIE-G-78-0213)

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-21.

Breland, H. M., & Braucher, J. L. Measuring writing ability. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Breland, H., & Ragosa, D. Validating placement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, N.J.: Educational Testing Service, 1976.

Cooper, C., Cherry, R., Gerber, R., Fleischer, S., Copley, B., & Sartisky, M. Writing abilities of regularly-admitted freshmen at SUNY/Buffalo. University Learning Center and Graduate Program in English Education, Department of Learning and Department of English, State University of New York, Buffalo, 1979.

Godshalk, F. I., Swineford, F., & Coffman, W. E. The measurement of writing ability. New York: College Entrance Examination Board, 1966.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Pitts, M. The relationship of classroom instructional characteristics and writing in the descriptive/narrative mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Popham, W. J. Domain-referenced strategies. In R. A. Berk (Ed.), Criterion-referenced measurement. Johns Hopkins University Press, 1980.

Powills, J. A., Bowers, R., & Conlan, G. Holistic essay scoring: An application of the model for the evaluation of writing ability and the measurement of growth in writing ability over time. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Quellmalz, E. Assessing writing proficiency: Designing integrated multi-level information systems. Paper presented at the annual meeting of the National Reading Conference, San Diego, CA, 1980.

Quellmalz, E. Defining writing domains: Effects of discourse and response mode. Interim report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (Grant No. OB-NIE-G-78-0213)

Quellmalz, E. Domain-referenced specifications for writing proficiency. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Quellmalz, E., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979. (OB-NIE-G-78-0213)

Smith, L. An assessment of writing needs of undergraduates in the life sciences and social sciences divisions at UCLA. Unpublished thesis, University of California, Los Angeles, 1975.

Spooner-Smith, L. Investigation of writing assessment strategies. Report to the National Institute of Education, November, 1978. (Grant No. OB-NIE-6-78-0213 to the UCLA Center for the Study of Evaluation)

Winters, L. The effects of differing response criteria on the assessment of writing competence. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1978. (Grant No. OB-NIE-G-78-0213)